<Bioinformatics Center>Bio-knowledge Engineering

Bioinformatics Center
– Bio-knowledge Engineering –

http://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
NGUYEN, Canh Hao
(D Knowledge Science)

Assist Prof
YAMADA, Makoto
(D Statistical Science)

Program-Specific Res
WIMALAWARNE, Kishan
(D Eng)

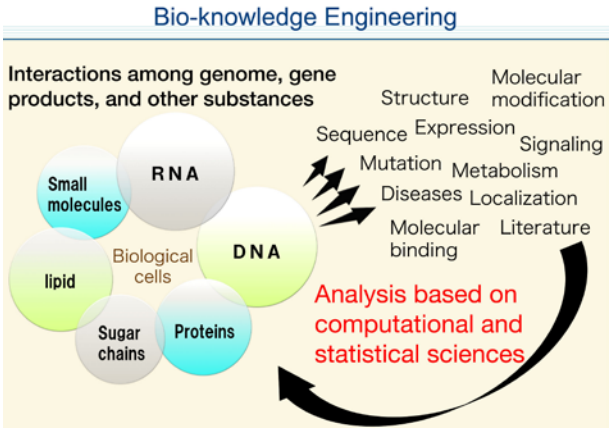
Students

TOHZAKI, Yudai (M1)

KIRIHATA, Tatsuro (UG)

Scope of Research

We are interested in graphs and networks in biology, chemistry, and medical sciences, including metabolic networks, protein-protein interactions and chemical compounds. We have developed original techniques in machine learning and data mining for analyzing these graphs and networks, occasionally combining with table-format datasets, such as gene expression and chemical properties. We have applied the techniques developed to real data to demonstrate the performance of the methods and find new scientific insights.



KEYWORDS

Bioinformatics Computational Genomics Data Mining
Machine Learning Systems Biology

Selected Publications

Gao, J.; Yamada, M.; Kaski, S.; Mamitsuka, H.; Zhu, S., A Robust Convex Formulations for Ensemble Clustering, *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, 1476-1482 (2016).

Peng, S.; You, R.; Wang, H.; Zhai, C.; Mamitsuka, H.; Zhu, S., DeepMeSH: Deep Semantic Representation for Improving Large-scale MeSH Indexing, *Bioinformatics 32 (12) (Proceedings of the 24th International Conference on Intelligent Systems for Molecular Biology (ISMB 2016))*, i70-i79 (2016).

Yuan, Q.-J.; Gao, J.; Wu, D.; Zhang, S.; Mamitsuka, H.; Zhu, S., DrugE-Rank: Improving Drug-Target Interaction Prediction of New Candidate Drugs or Targets by Ensemble Learning to Rank, *Bioinformatics 32 (12) (Proceedings of the 24th International Conference on Intelligent Systems for Molecular Biology (ISMB 2016))*, i18-i27 (2016).

Mohamed, A.; Nguyen, C. H.; Mamitsuka, H., NMRPro: An Integrated Web Component for Interactive Processing and Visualization of NMR Spectra, *Bioinformatics*, **32,** 2067-2068 (2016).

Shinkai-Ouchi, F.; Koyama, S.; Ono, Y.; Hata, S.; Ojima, K.; Shindo, M.; duVerle, D.; Kitamura, F.; Doi, N.; Takigawa, I.; Mamitsuka, H.; Sorimachi, H., Predictions of Cleavability of Calpain Proteolysis by Quantitative Structure-Activity Relationship Analysis Using Newly Determined Cleavage Sites and Catalytic Efficiencies of an Oligopeptide Array, *Mol. Cell. Proteomics*, **15,** 1262-1280 (2016).

Nguyen, C. H.; Mamitsuka, H., New Resistance Distances with Global Information on Large Graphs, *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2016) (JMLR Workshop and Conference Proceedings,51)*, 639-647 (2016).

Ultra High-Dimensional Nonlinear Feature Selection for Big Biological Data

Life sciences are going through a revolution thanks to the possibility to collect and learn from massive biological data. Efficient processing of such "big data" is extremely important for many medical and biological applications, including disease classification, biomarker discovery, and drug development. The complexity of biological data is dramatically increasing due to improvements in measuring devices such as next-generation sequencers, microarrays and mass spectrometers. As a result, we must deal with data that includes many observations (hundreds to tens of thousands) and even larger numbers of features (thousands to millions). Machine learning algorithms are charged with learning patterns and extracting actionable information from biological data. These techniques have been used successfully in various analytical tasks, such as genome-wide association studies and gene selection.

However, the scale and complexity of big biological data pose new challenges to existing machine learning algorithms. There is a trade-off between scalability and complexity: linear methods scale better to large data, but cannot model complex patterns. Nonlinear models can handle complex relationships in the data but are not scalable to the size of current datasets. In particular, learning nonlinear models requires a number of observations that grows exponentially with the number of features. Biological data generated by modern technology has as many as millions of features, making the learning of nonlinear models unfeasible with existing techniques. To make matters worse, current nonlinear approaches cannot take advantage of distributed computing platforms.

A promising approach to make nonlinear analysis of big biological data computationally tractable is to reduce the number of features. This method is called feature selection. Biological data is often represented by matrices where rows denote features and columns denote observations. Feature selection aims to identify a subset of features (rows) to be preserved, while eliminating all others.

Here we propose a novel feature selection framework for big biological data that makes it possible for the first time to identify very few relevant, non-redundant features among millions. The proposed method is based on two components: Least Angle Regression (LARS), an efficient feature selection method, and the Hilbert-Schmidt Independence Criterion (HSIC), which enables the selection of features that are non-linearly related. These properties are combined to obtain a method that can exploit nonlinear feature dependencies efficiently, and furthermore enables distributed implementation on commodity cloud computing platforms. We name our algorithm Least Angle Nonlinear Distributed (LAND) feature selection. Through experiment on a large and high-dimensional dataset (million features with tens of thousand samples), we show that our approach can find a set of independent features without losing classification accuracy (Figure 1 A and B). Moreover, the result is achieved in hours of cluster computing time (Figure1 C). The selected features are relevant and non-redundant, making it possible to obtain accurate and interpretable models that can be run on a laptop computer.

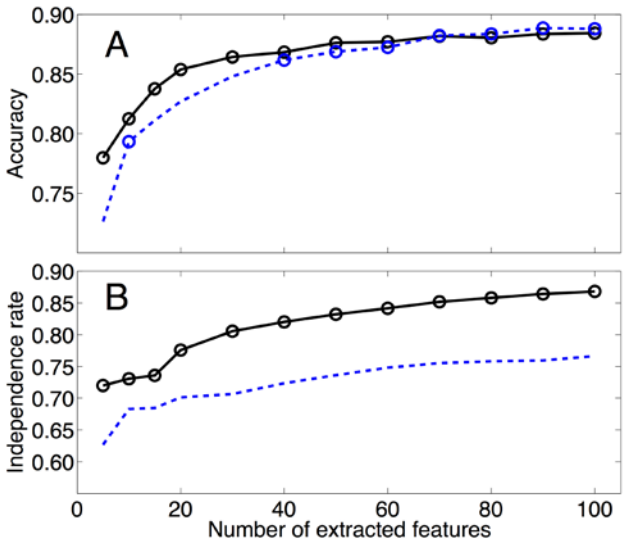


Figure 1. Results for the enzyme dataset. Circles indicate best accuracy/independence rate according to t-tests ($p<0.05$). Differences in accuracy are significant for m=5,20,30,40. (B)~Independence rate vs. m: all differences are significant.

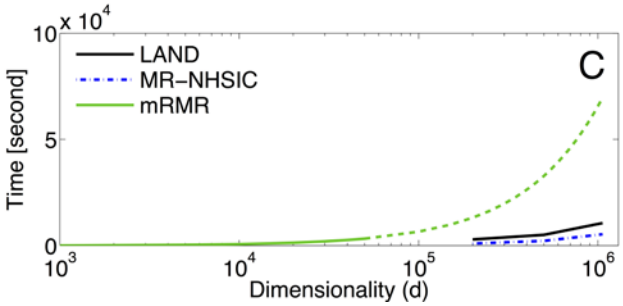


Figure 2. Computation time of feature selection algorithms. We ran the algorithms on reduced-dimensionality datasets. We estimated the mRMR runtime for d> 50,000, given that it scales linearly with d (dotted line).